Scale Adaptive Fusion Network for RGB-D Salient Object Detection

Yuqiu Kong[1], Yushuo Zheng[2], Cuili Yao[1], Yang Liu[1], and He Wang[3]

[1] School of Innovation and Entrepreneurship, Dalian University of Technology, Dalian, China
[2] School of Mechanical Engineering, Dalian University of Technology, Dalian, China
[3] School of Computer Science and Technology, Dalian University of Technology, Dalian, China
{yqkong,yaocuili1984,ly}@dlut.edu.cn
{2256357628,mrs_wangher}@mail.dlut.edu.cn

Abstract. RGB-D Salient Object Detection (SOD) is a fundamental problem in the field of computer vision and relies heavily on multi-modal interaction between the RGB and depth information. However, most existing approaches adopt the same fusion module to integrate RGB and depth features in multiple scales of the networks, without distinguishing the unique attributes of different layers, *e.g.*, the geometric information in the shallower scales, the structural features in the middle scales, and the semantic cues in the deeper scales. In this work, we propose a Scale Adaptive Fusion Network (SAFNet) for RGB-D SOD which employs scale adaptive modules to fuse the RGB-D features. Specifically, for the shallow scale, we conduct the early fusion strategy by mapping the 2D RGB-D images to a 3D point cloud and learning a unified representation of the geometric information in the 3D space. For the middle scale, we model the structural features from multi-modalities by exploring spatial contrast information from the depth space. For the deep scale, we design a depth-aware channel-wise attention module to enhance the semantic representation of the two modalities. Extensive experiments demonstrate the superiority of the scale adaptive fusion strategy adopted by our method. The proposed SAFNet achieves favourable performance against state-of-the-art algorithms on six large-scale benchmarks.

Keywords: RGB-D salient object detection · Multi-modal analysis and understanding · Multi-modal fusion strategy

1 Introduction

Salient object detection, aiming to locate and recognize the most attractive regions in the scene, has received wide research interest in recent years. As an effective pre-processing method, it has been applied to various computer vision tasks, such as scene classification [33], visual tracking [25], image editing [48], *etc.* Although RGB SOD methods [37,38,39] achieve satisfactory results in natural scenes, their performances are limited when the scenes are complicated or the

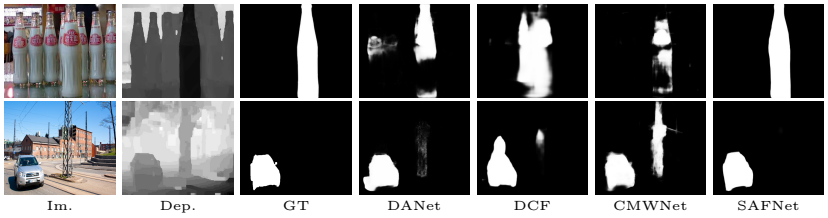

Fig. 1. Saliency prediction results from different fusion methods. From left to right: the input RGB and depth images, the ground truth images, saliency maps of the state-of-the-art detectors, DANet [45], DCF [14], CMWNet [20], and our proposed SAFNet. DANet and DCF employ the same fusion methods in multiple scales. CMWNet and SAFNet design adaptive fusion modules for different scales. Compared with them, the SAFNet can generate more accurate predictions by exploring complementary information from cross-modalities with effective fusion strategies.

appearance of targets is not dominant in the RGB space. With the development of depth cameras, researchers can learn geometric and location information from the depth image which is complementary to the RGB image. It helps identify salient objects from distractors and leads to discriminative SOD models even in very cluttered environments (see Fig. 1).

Considering that there is a large gap between the distributions of RGB and depth data, existing RGB-D SOD algorithms usually focus on exploring effective fusion strategies to model the complementary information between the two modalities. These fusion strategies can be classified into early fusion [12,45], mid-level fusion [43,3,20,41,14,19,15], and late fusion [10,5]. Although these fusion strategies have improved the performance of saliency models, some issues still exist to be considered. First, the early fusion strategy assembles the RGB-D images (*e.g.*, concatenation) and then feeds them into feature extractors. However, the RGB and depth images incorporate asynchronous information. The simple concatenation operation will eliminate distinctive features provided by the two modalities. In addition, feature extractors (*e.g.*, VGG [34], ResNet [11]) are usually pre-trained on the RGB-based benchmarks, they are insufficient to learn both appearance and geometric features from the combined RGB-D data. Second, mid-level fusion strategies are the most important operations to integrate the cross-modal features from the RGB and depth images. However, most existing algorithms design and employ the same fusion operation in different scales of the network, *e.g.*, the DANet [45] and DCF [14] in Fig. 1. They ignore unique attributes of features in multiple scales, such as the appearance and geometric information in the shallower scales, the structure cues in the middle scales, and the high-level semantic feature in the deeper scales. Despite CMWNet [20] considering the diversity of multi-scale features, it suffers from the inferior representation ability of fusion modules. Therefore, these methods demonstrate a limited capacity to explore discriminative cross-modal features from different levels of the network and lead to sub-optimal performance of the final prediction. Such as the visual examples in Fig. 1, DANet, DCF, and CMWNet fail to

capture the true targets (the 1st row) and wrongly respond to the distractive regions in the image background (the 2nd row).

To address the above issues, a Scale Adaptive Fusion Network (SAFNet) is proposed for RGB-D SOD. We conduct in-depth studies of both early and mid-level fusion strategies and encourage the multi-scale interactions of cross-modalities. The SAFNet is a two-stream network and adaptively integrates RGB-D features in shallow, middle, and deep scales by a cross-modal fusion encoder. For the first issue, instead of early fusing the RGB-D images in the 2D space, we project them into the 3D space and represent them as the point cloud data. By learning the point cloud representation in the 3D space, we can explicitly model the pixel-wise affinity and explore the appearance and geometric information. For the second issue, we elaborate on the fusion module for each scale according to its characteristic to fully exploit the complementary information from the two modalities. Specifically, to combine cross-modal features from the shallower scale without eliminating the geometric information, we propose a Point Cloud based Fusion (PCF) method. It employs the PointNet++ [31] to learn point-wise representation in the 3D point cloud space, so that the network can explore detailed cues around the neighbourhood. For the middle scale features which abstract structure information from the shallower scale, we propose a Spatial Contrast Refinement (SCR) module to refine the integrated RGB-D features. By exploiting spatial contrastive information from the depth data, the network can learn more discriminative representations from the RGB-D features to better distinguish the targets from the background. To enhance the representation ability of the semantic feature on the deep scale, we design a Depth-aware Channel-wise Attention (DCA) module to associate the synchronous feature from the cross-modalities.

The main contributions of our work are three folds: (1) (1) We propose a SAFNet for the RGB-D SOD. Effective early fusion and mid-level fusion strategies are studied in this work. We focus on designing scale adaptive fusion modules to sufficiently explore the complementary information from the cross-modalities. (2) To fuse the multi-modal features, in the shallow scale, a PCF module is proposed to integrate the appearance and geometric information in the 3D point cloud space; in the middle scale, an SCR module is designed to model the structural information in the scene; in the deep scale, a DCA module is adopted to enhance the representation ability of semantic cues. (3) Quantitative and qualitative on six large-scale datasets demonstrate the effectiveness of the proposed fusion strategies, and our method achieves favourable performance compared with the state-of-the-art algorithms.

2 Related Work

2.1 RGB-D SOD Methods

With the development of convolutional neural networks, the performances of RGB-D SOD methods have gained significant improvement compared to traditional hand-crafted based methods [18,4,32]. The recent RGB-D SOD models

can be roughly categorized into single-stream methods [47,45] and two-stream methods [30,43,35,46]. The work [47] uses a small encoder to extract prior information from the depth image to enhance the robustness of the main RGB stream. Zhao *et al.* [45] design a lightweight single stream network that employs the depth map to guide the early and mid-level fusion between the RGB-D images. Although single stream networks can save computation costs and lead to real-time models, they are limited in representing geometric features of the depth image and integrating the multi-modal information. In contrast, two-stream networks separately extract features of RGB-D images and fuse them in different scales. For example, Zhang *et al.* [43] propose an asymmetric two-stream network, in which a flow ladder module is designed to explore the global and local cues of RGB image, and a depth attention module is designed to improve the discriminative ability of the fused RGB-D feature. The work [46] trains a specificity-preserving network to explore modality-specific attributes and shared information of RGB-D images. In addition, various learning strategies are exploited to enhance the interaction between the multi-modalities. In work [42], the mutual learning strategy is applied on each scale of the two-stream network to minimize the distance of the representations of RGB-D modalities. Ji *et al.* [14] first calibrate the depth data using a learning strategy and then fuse the RGB-D features with a cross reference module. The work [44] employs the self-supervised representation learning method to pre-train the network using only a few unlabelled RGB-D datasets, thereby learning good initialization parameters for the downstream saliency prediction task.

2.2 Cross-Modality Fusion Models for RGB-D SOD

The main concerns of RGB-D SOD methods are how to 1) integrate complementary information between the RGB-D modalities and 2) enhance the combination of the consistent semantic features from RGB-D images. To this end, recent RGB-D SOD methods design early fusion and mid-level fusion strategies to achieve the in-depth fusion between the two modalities. The early fusion based methods [24,45] usually concatenate the RGB-D images on the channel dimension. However, they ignore the distribution difference between the multi-modalities. Therefore, researchers make efforts to design effective mid-level fusion methods. In [43,45], attention based modules are proposed to select informative depth cues and provide guidance for the interaction of the two modalities. A cross reference module is proposed in [14] to fuse complementary features from RGB-D images. Zhou *et al.* [46] employ a cross-enhanced integration module to fuse RGB-D features in different scales of the encoder, and design a multi-modal feature aggregation module to gather modality-specific features in the decoder. Sun *et al.* [35] utilize a depth-sensitive attention module to achieve automatic multi-modal multi-scale integration for RGB-D SOD.

Although the above methods design effective modules or training strategies for the cross-modal fusion of two modalities, they usually apply the same fusion modules on different scales and thus ignore the distribution difference of features in multiple scales. In this work, we study the unique attribute of each scale and

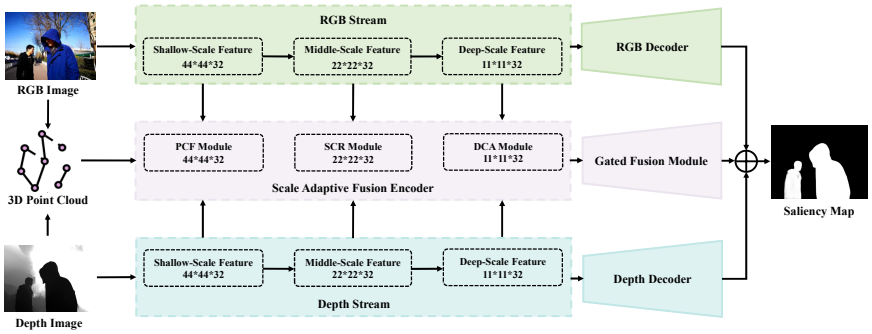

Fig. 2. Architecture of the proposed SAFNet.

design adaptive feature fusion methods for them, which leads to more satisfactory saliency detection results.

3 Algorithm

In this section, we first overview the architecture of the proposed SAFNet and then elaborate on each component of the network. The structure of the SAFNet is shown in Fig. 2.

3.1 Overview

Most existing methods for RGB-D SOD employ the same cross-modal fusion operation in different scales of the network, thus inevitably ignoring the distinctive attributes of multi-scale features. In this work, we propose a two-stream network, SAFNet, to adaptively integrate multi-modality features in different scales with customized fusion modules.

Given the paired RGB image I_c and depth image I_d, we adopt a two-stream network to extract the multi-scale features of the two modalities separately. Following the work [14], for each stream, we employ the architecture of work [37] as the backbone to generate hierarchical features. Its detailed structure is shown in Fig. 3. The three-level features before the second partial decoder are used as the shallow, middle, and deep scale features, respectively, which are denoted as $\{f_c^i\}_{i=1}^3$ for the RGB image, and $\{f_d^i\}_{i=1}^3$ for the depth image. It is followed by a scale adaptive fusion encoder which integrates the hierarchical features from cross-modalities according to their scale attributes. Specifically, for the shallow scale, we propose a Point Cloud Fusion (PCF) module which utilizes the PointNet++ method as the feature extractor $F_{shallow}(\cdot,\cdot)$ to learn the feature representation in the 3D point cloud space, $f_{shallow} = F_{shallow}(f_c^1, f_d^1)$. For the middle scale, we design a Spatial Contrast Refinement (SCR) module to refine the integrated multi-modal feature by exploring the spatial contrastive information. By this means, we can model the structural features in the scene. This process is denoted as $f_{middle} = F_{middle}(f_c^2, f_d^2)$. For the deep scale, a Depth-aware

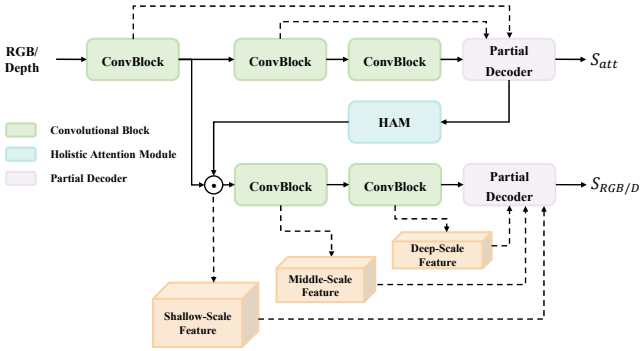

Fig. 3. Architecture of the RGB stream and depth stream. The implementations of the convolutional block, the holistic attention module, and the partial decoder follow the work [37].

Channel-wise Attention (DCA) module is adopted to enhance the representation ability of semantic cues from the RGB-D modalities, $f_{deep} = F_{deep}(f_c^3, f_d^3)$. The hierarchical integrated features, $f_{shallow}, f_{middle}$, and f_{deep} are then fed into the Gated Fusion Module (GFM) to generate a saliency map S_{fusion}. The final saliency map can be obtained by averaging S_{fusion} and the saliency predictions of the RGB and depth streams, S_{RGB} and S_D. The architecture of the fusion stream is illustrated in Fig. 4. Note that each partial decoder in the network generates a saliency map, which is supervised by the ground truth image.

3.2 Scale Adaptive Fusion Encoder

In this section, we present the architecture of the scale adaptive fusion encoder and elaborate on fusion modules at shallow, middle, and deep scales.

Point Cloud Fusion Module for Shallow Scale. Since features from RGB and depth modalities in the shallow scales, f_c^1 and $f_d^1 \in \mathcal{R}^{H_1 \times W_1 \times C_1}$, incorporate valuable cues of appearance and geometric information, it is more suitable to integrate them in the 3D space. We first project the input RGB-D images into the point cloud representation by transforming the 2D image coordinate (x, y) to the world coordinate system $(x^{'}, y^{'}, z^{'})$,

$$\begin{bmatrix} x^{'} \\ y^{'} \\ z^{'} \end{bmatrix} = \frac{I_d(x,y)}{s} \begin{bmatrix} \frac{1}{f_x} & 0 & 0 \\ 0 & \frac{1}{f_y} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}, \tag{1}$$

where $I_d(x, y)$ is the depth value in position (x, y), f_x and f_y are the focal length parameters, and s is the scaling factor of the camera. The point-wise feature in the point cloud space can be initially represented by concatenating

the shallow scale features of RGB and depth images along the channel dimension, $f_{pcd} = [f_c^1, f_d^1]$.

To learn the point-wise representation, we employ PointNet++ [31] to capture fine-grained details around the neighbourhood at multiple levels. The PointNet++ is an encoder-decoder architecture. Fed with the initial point cloud feature matrix f_{pcd} with the size $N \times (d + C)$, where $N(= H_1 \times W_1), d$, and C are the number of points, the dimension of coordinates, and the dimension of point-wise feature respectively, the encoder learn the representation hierarchically by a set of *set abstraction levels* (3 in our work), consisting of a sampling layer, a grouping layer, and a PointNet [9] layer. In each set abstraction level, the sampling layer chooses $N^{'}$ $(N^{'} < N)$ points as centroids that best cover the entire point cloud with appropriate receptive fields. Then the grouping layer gathers the neighbour points of each centroid into a group with a ball query strategy. The output is a group of points with the size $N^{'} \times K \times (d + C)$, where K is the upper limit number of points in the neighbourhood of the centroids. Finally, the PointNet layer encodes the feature of each group to learn the local pattern. The output feature matrix is of size $N^{'} \times (d + C^{'})$, where $C^{'}$ is the dimension of the feature. By sub-sampling the point cloud with the hierarchical set abstraction levels, the local context of the point cloud is captured at multiple scales. To learn the feature of each original point in the point cloud, the decoder adopts a set of *feature propagation levels*, skip links, and Multilayer Perceptrons (MLP) to propagate point features output by the encoder to all original points in a hierarchical manner. The output feature $\hat{f}_{shallow}$ has the size of $N \times C$ and then is reshaped to the size of $H_1 \times W_1 \times C$.

3D-2D Normalization. Since the fused feature in the shallow scale is in the 3D space, to seamlessly integrate it with subsequent features, we employ a normalization operation to project it to the 2D space to alleviate the difference in distributions.

$$f_{shallow} = InstanceNorm(RFB(\hat{f}_{shallow})), \tag{2}$$

where $InstanceNorm(\cdot)$ is the instance normalization layer [36], and $RFB(\cdot)$ is the Receptive Field Block (RFB) in work [37]. Note that if we replace the instance normalization layer with the batch normalization layer [13], the operation in Eq. 2 will degrade to a flat operation in the network. In Sec. 4.4, we experimentally verify the effectiveness of the instance normalization operation.

Spatial Contrast Refinement Module for Middle Scale. Mid-level layers of the RGB and depth streams encode the features of shallower scales and learn the structure information of the scene. To integrate middle-scale features of multiple modalities and explore the mild-level structure cues, we introduce a Graph Neural Network (GNN) that refines the fused multi-modal features according to the spatial contrastive relationship with other regions. We first wrap the middle-scale features of RGB-D images, f_c^2 and f_d^2, with a convolutional layer and then concatenate them in the channel dimension,

$$f_{mid} = Concat(Conv_{\theta_c}(f_c^2), Conv_{\theta_d)}(f_d^2), \tag{3}$$

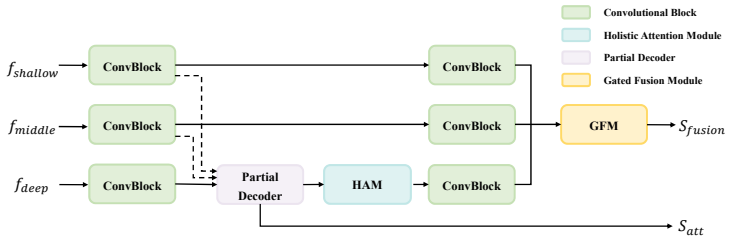

Fig. 4. Architecture of the fusion stream, which takes the multi-scale integrated features ($f_{shallow}$, f_{middle}, and f_{deep}) as input, and outputs a saliency map S_{fusion}.

where $Concat(\cdot,\cdot)$ is the concatenation operation, $Conv_{\theta_c}(f)$ and $Conv_{\theta_d}(f)$ are the convolution operations with parameters θ_c and θ_d, respectively. Then a fully-connected graph $G = (V, E)$ is constructed over the feature f_{mid}, where V is the node set and E is the edge set. Specifically, each pixel in f_{mid} is regarded as a node $n_i \in V(i = 1, 2, \cdots K)$, and K is the number of pixels in f_{mid}. The edge between node n_i and n_j is denoted as e_{ij}, and is weighted by the distance in the 3D space between the nodes,

$$w_{ij} = exp(-|(x_i, y_i, d_i) - (x_j, y_j, d_j)|), \tag{4}$$

where (x, y) is the spatial coordinate of the node, d is the depth value. In addition, we depict the semantic affinity feature a_{ij} between each pair of nodes,

$$a_{ij} = MLP(Concat(f_{mid,i}, f_{mid,j}, f_g)), \tag{5}$$

where $f_{mid,i}, f_{mid,j}$ are feature vectors of nodes n_i and n_j, MLP is the multilayer perceptron, and f_g is the global semantic feature by applying the Global Average Pooling (GAP) on f_{mid}. Based on graph G, we can update the feature of each node by a graph convolutional layer,

$$\hat{f}_{mid,i} = MLP(Concat(\sum_{j \in \mathcal{N}(n_i)} w_{ij} a_{ij}, f_{mid,i}, f_{mid,j}), \tag{6}$$

where the set $\mathcal{N}(n_i)$ contains neighbours of the node n_i. We denote the output feature map of the GNN as $\hat{f}_{mid}$ by spatially arranging the feature $\{\hat{f}_{mid,i}\}_{i=i}^{K}$. The weight w_{ij} and the semantic feature a_{ij} indicate the spatial affinity and semantic correlation between the nodes n_i and n_j, respectively. Therefore, the GNN based refinement operation encourages closer regions, in both Euclidean and appearance space, to contribute more to the refined features. By this means, relevant pixels with the same saliency labels tend to be gathered together, leading to more accurate prediction results.

Depth-Aware Channel-Wise Attention Module for Deep Scale. We adopt a DCA module to make an effective alignment between the deep-scale

features of multi-modalities. Compared to the deep-scale feature of the depth image, the high-level semantic information in the RGB feature is more crucial for saliency detection. The spatial-level integration (*e.g.*, spatial-wise attention) may introduce a negative effect on the final prediction. Therefore, we learn the channel-wise attention vector from the deep-scale depth feature and use it to highlight the significant dimensions of the deep-scale RGB feature.

First, the deep-scale depth feature map $f_d^3 \in \mathcal{R}^{H_3 \times W_3 \times C_3}$ is encoded by a convolutional layer and is transformed into a 1-channel feature map. Then the feature map is reshaped into the size of $N_3 \times 1$, where $N_3 = H_3 \times W_3$, and a softmax layer is applied to it to generate a pixel-wise attention vector f_{d_att} which indicates the spatial significance of the depth feature. Second, the deep-scale RGB feature map f_c^3 is reshaped to the size of $N_3 \times C_3$, and is multiplied with the attention vector,

$$a_c = [R(f_c^3)]^T \times f_{d\_att}, \tag{7}$$

where $R(\cdot)$ is the reshape operation, and a_c is of the size $C_3 \times 1$. The above implementation learns the correlation between the depth and RGB modalities and adaptively integrates the appearance feature according to the spatial significance. We then utilize the channel-wise attention $a_c \in \mathcal{R}^{C_3 \times 1}$ to highlight the important dimension of the RGB feature map f_d^3,

$$f_{deep} = tile(a_c) \odot f_d^3, \tag{8}$$

in which the function $tile(\cdot)$ tiles the channel-wise attention vector to the size of f_d^3, and $\odot$ is the element-wise multiplication. As a result, the output feature map f_{deep} is equipped with a powerful representation ability of high-level semantic information.

3.3 Saliency Decoder

The decoders of the RGB and depth streams are based on the structure of the work [37], and output saliency maps S_{RGB} and S_D, respectively. In this section, we mainly elaborate on the saliency decoder of the fusion stream, which takes advantage of the multi-modal fusion features in different scales and generates a saliency map S_{fusion}.

First, the fusion features $f_{shallow}$, f_{mid}, and f_{deep} are wrapped by a Receptive Field Block (RFB) [23] to explore global context information, respectively. The output features are decoded as $\hat{f}_{shallow}$, $\hat{f}_{mid}$, and $\hat{f}_{deep}$. Then a Gated Fusion Module (GFM) is designed to integrate the multi-scale features successively,

$$\begin{cases} f_{sm} &= Gate(Concat(\hat{f}_{shallow}, \hat{f}_{mid})) \odot \hat{f}_{shallow} + \hat{f}_{mid}, \\ f_{md} &= Gate(Concat(\hat{f}_{mid}, \hat{f}_{deep})) \odot \hat{f}_{mid} + \hat{f}_{deep}, \\ f_{smd} &= Gate(Concat(f_{sm}, f_{md})) \odot f_{sm} + f_{md}, \end{cases} \tag{9}$$

where $Gate(\cdot)$ is the gated function and formulated as a sequence of a 3×3 convolutional layer, a batch normalization layer, and a sigmoid layer. The gated

function learns spatial-wise attention from features of multiple scales, which are then adaptively integrated by the guidance of attention. The output feature map f_{smd} incorporates effective information from different levels, and based on this, a prediction result S_{fusion} is generated using a 3×3 convolutional layer.

The final saliency map is obtained by averaging the saliency maps output by the three decoders, namely $S_{final} = (S_{RGB} + S_D + S_{fusion})/3$.

3.4 Loss Function

The proposed SAFNet is trained in an end-to-end manner. The total loss function $\mathcal{L}_{total}$ is a summation of the losses from the RGB stream, the depth stream, and the fusion stream, denoted as $\mathcal{L}_{RGB}, \mathcal{L}_D$, and $\mathcal{L}_{fusion}$, respectively,

$$\mathcal{L}_{total} = \mathcal{L}_{RGB} + \mathcal{L}_D + \mathcal{L}_{fusion}. \tag{10}$$

Each loss is defined as the binary cross-entropy loss between the predicted saliency map and the ground truth image,

$$\mathcal{L}_k(S_k, G) = G \log S_k + (1 - G) \log(1 - S_k), k \in \{RGB, D, fusion\} \tag{11}$$

where S_k is the saliency map and G is the ground truth image.

4 Experiments

4.1 Implementation Details

All experiments are implemented on the PyTorch platform with a single 2080Ti GPU. The backbone of each stream in the encoder is based on the ResNet-50 [11] and is initialized by the pre-trained parameters in ImageNet [17]. All input RGB and depth images are resized to 352×352. We employ the common-used data augmentation methods incorporating randomly horizontal flipping, cropping, and rotating. In the process of point cloud representation learning , both the focal lengths f_x and f_y are set as 44 (the same as the spatial size of the shallow-scale feature maps), and the scaling factor s is 255.0. The number of centroids in the 3 set abstraction levels is set as 2048, 1024, and 512, respectively. The number of channels of feature maps (C, C', C^1, C^2, and C^3) is 32. In the training stage, the batch size is set as 8. We employ the Stochastic Gradient Descent (SGD) algorithm with momentum 0.9 to optimize the objective function. The learning rate is 0.0005. The network converges within 200 epochs.

4.2 Datasets and Metrics

Datasets. We evaluate the performance of the proposed SAFNet and the compared methods on six public RGB-D SOD benchmarks, including DUT-D [29] (1200 image pairs), NLPR [28] (1000 image pairs), NJUD [16] (1985 stereo image pairs), STERE [26] (1000 image pairs), SIP [8] (929 image pairs), and LFSD [21]

Table 1. The maximum F-measure, S-measure, E-measure, and MAE of the evaluated saliency models on six data sets. The top three scores of each method are marked as red, green, and blue, respectively.

	Metric	S2MA CVPR20	UCNet CVPR20	HDFNet ECCV20	DANet ECCV20	PGAR ECCV20	CMWNet ECCV20	ATSA ECCV20	SPNet ICCV21	D3Net TNNLS21	DSA2F CVPR21	DCF CVPR21	HAINet TIP21	CDNet TIP21	CDINet ACMM21	SSP AAAI22	SAFNet Ours
DUT-D	F_{max}	0.909	-	0.926	0.911	0.938	0.905	0.936	-	-	0.938	0.941	0.932	0.944	0.934	0.947	0.950
	S_m	0.903	-	0.905	0.889	0.920	0.887	0.916	-	-	0.921	0.924	0.909	0.927	0.927	0.929	0.931
	E_m	0.921	-	0.938	0.929	0.950	0.922	0.954	-	-	0.956	0.957	0.939	0.957	0.956	0.958	0.962
	MAE	0.044	-	0.040	0.047	0.035	0.056	0.033	-	-	0.031	0.030	0.038	0.031	0.029	0.029	0.026
NLPR	F_{max}	0.910	0.916	0.917	0.908	0.925	0.913	0.905	0.925	0.907	0.916	0.917	0.917	0.928	0.916	0.923	0.934
	S_m	0.915	0.920	0.916	0.908	0.930	0.917	0.911	0.927	0.912	0.918	0.921	0.921	0.927	0.927	0.922	0.93
	e_m	0.942	0.955	0.948	0.945	0.954	0.941	0.947	0.959	0.945	0.952	0.956	0.952	0.955	0.960	0.960	0.965
	MAE	0.030	0.025	0.027	0.031	0.025	0.029	0.028	0.021	0.030	0.024	0.023	0.025	0.025	0.024	0.025	0.02
NJUD	F_{max}	0.899	0.908	0.924	0.905	0.918	0.913	0.904	0.935	0.910	0.917	0.917	0.920	0.919	0.921	0.923	0.929
	S_m	0.894	0.897	0.911	0.897	0.909	0.903	0.887	0.924	0.900	0.904	0.903	0.909	0.913	0.918	0.909	0.915
	E_m	0.917	0.934	0.934	0.926	0.935	0.923	0.926	0.953	0.916	0.937	0.941	0.931	0.940	0.951	0.939	0.947
	MAE	0.053	0.043	0.037	0.046	0.042	0.046	0.047	0.029	0.046	0.039	0.038	0.038	0.038	0.036	0.038	0.033
STERE	F_{max}	0.895	0.908	0.918	0.897	0.911	0.911	0.911	0.915	0.904	0.910	0.915	0.919	0.908	-	0.914	0.920
	S_m	0.890	0.903	0.906	0.892	0.907	0.905	0.896	0.907	0.899	0.897	0.905	0.909	0.903	-	0.885	0.907
	E_m	0.926	0.942	0.937	0.927	0.937	0.930	0.942	0.942	0.924	0.942	0.943	0.938	0.938	-	0.929	0.944
	MAE	0.051	0.039	0.039	0.048	0.041	0.043	0.038	0.037	0.046	0.039	0.037	0.038	0.041	-	0.047	0.036
SIP	F_{max}	0.891	0.896	0.904	0.901	0.893	0.890	0.885	0.916	0.881	0.891	0.900	0.916	0.888	0.884	0.895	0.914
	S_m	0.872	0.875	0.878	0.878	0.876	0.867	0.852	0.894	0.860	0.862	0.873	0.886	0.862	0.875	0.868	0.886
	E_m	0.913	0.918	0.921	0.917	0.912	0.909	0.899	0.930	0.902	0.911	0.921	0.925	0.905	0.915	0.910	0.927
	MAE	0.057	0.051	0.050	0.054	0.055	0.062	0.064	0.043	0.063	0.057	0.052	0.048	0.060	0.055	0.058	0.047
LFSD	F_{max}	0.862	0.878	0.882	0.871	0.834	0.900	0.883	0.881	0.840	0.903	0.867	0.880	0.898	0.890	0.870	0.901
	S_m	0.837	0.865	0.855	0.849	0.816	0.876	0.855	0.854	0.825	0.883	0.841	0.859	0.878	0.870	0.853	0.872
	E_m	0.863	0.906	0.879	0.881	0.870	0.908	0.897	0.897	0.863	0.923	0.883	0.895	0.912	0.914	0.891	0.914
	MAE	0.095	0.067	0.078	0.079	0.091	0.066	0.071	0.071	0.095	0.055	0.075	0.072	0.061	0.063	0.075	0.061

(100 image pairs). As the training settings of most methods, we select 800 samples from the DUT-D, 1485 samples from the NJUD, and 700 samples from the NLPR as the training set. The rest samples are treated as test sets.

Metrics. Four metrics are employed to evaluate the performance of saliency models, including the maximum F-measure [1] (F_{max}), S-measure [6] (S_m), E-measure [7] (E_m), and Mean Absolute Error (MAE) [2]. F-measure values comprehensively consider the precision and recall of saliency models. S-measure values measure the structural similarity between saliency maps and ground truth images. E-measure values capture pixel-level matching and image-level statistics information. MAE values depict errors in predicted saliency maps.

4.3 Comparison with State-of-the-arts

We compare the proposed SAFNet with 15 state-of-the-art algorithms, including S2MA [22], UCNet [41], HDFNet [27], DANet [45], PGAR [3], CMWNet [20], ATSA [43], SPNet [46], D3Net [8], DSA2F [35], DCF [14], HAINet [19], CDNet [15], CDINet [40], and SSP [44]. For a fair comparison, the evaluated saliency maps are provided by their authors or generated by the public released codes.

Quantitative Evaluation. The quantitative performances of evaluated methods in terms of F-measure, S-measure, E-measure, and MAE are demonstrated in Tab. 1. It shows that the proposed SAFNet achieves satisfactory results and outperforms most state-of-the-art algorithms on six challenging datasets, indicating the generalization ability of our method.

Qualitative Evaluation. We illustrate the visual results of the evaluated methods in Fig. 4.2. It shows that our method can handle various challenging scenarios. For the first example which incorporates multiple objects with different appearances, our proposed SAFNet can capture all salient objects compared to

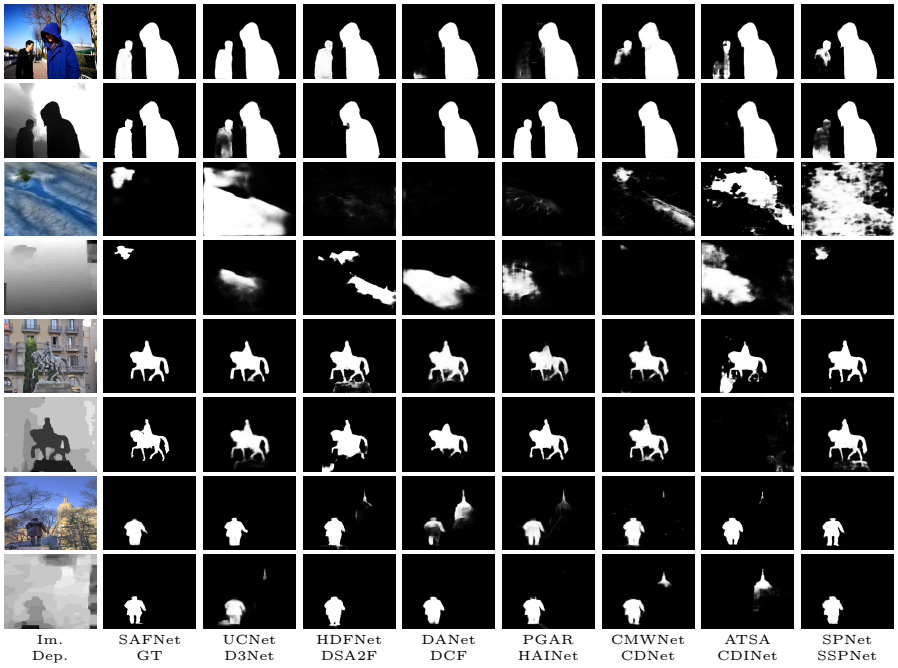

Fig. 5. Visual results of the proposed SAFNet and the compared methods.

many existing methods. In the second example, the target object is not salient in the color space and the size is small. The depth image provides valuable cues about the object and background. By properly integrating the complementary information of the RGB-D modalities, our method successfully distinguishes the salient object from the background. In contrast, most other methods fail to capture the target and wrongly respond to the background regions. In the third example, the appearance of the salient object is similar to the background. Many other methods present blurry predictions around the object boundary. Since the depth image shows obvious structural information, our method can segment the target accurately from the cluttered background. For the fourth example, many existing methods wrongly capture part of the building as saliency regions. In contrast, our method learns the contrast information in both color and depth space and recognizes accurate regions as saliency.

4.4 Ablation Studies

We conduct ablation studies to verify the effectiveness of the main components in the proposed SAFNet. We also compare the key modules of our method with the advanced fusion methods of existing algorithms. The rationality of the scale adaptive fusion strategy is validated.

Table 2. Ablation studies on each main component of the SAFNet.

Model	NLPR				NJDU				DUT-D			
	F_{max}	S_m	E_m	MAE	F_{max}	S_m	E_m	MAE	F_{max}	S_m	E_m	MAE
(a) RGB-D Baseline (B)	0.897	0.891	0.934	0.036	0.894	0.896	0.915	0.051	0.911	0.895	0.932	0.044
(b) B+PCF	0.917	0.916	0.955	0.024	0.918	0.906	0.942	0.038	0.939	0.917	0.952	0.032
(c) B+PCF+SCR	0.921	0.918	0.955	0.024	0.919	0.908	0.094	0.036	0.941	0.919	0.955	0.030
(d) B+PCF+SCR+DCA	0.928	0.926	0.962	0.021	0.924	0.911	0.946	0.035	0.946	0.927	0.959	0.028
(e) B+PCF(BN norm)+SCR+DCA+GFM	0.930	0.929	0.965	0.020	0.926	0.913	0.944	0.034	0.948	0.928	0.960	0.026
(f) B+PCF+SCR+DCA+GFM (SAFNet)	0.934	0.930	0.965	0.020	0.929	0.915	0.947	0.033	0.950	0.931	0.962	0.026

Table 3. Ablation studies on fusion strategies.

DCADCA	NLPR				NJDU				DUT-D			
	F_{max}	S_m	E_m	MAE	F_{max}	S_m	E_m	MAE	F_{max}	S_m	E_m	MAE
(a) RGB-D Baseline (B1)	0.903	0.904	0.942	0.029	0.902	0.896	0.927	0.043	0.924	0.904	0.940	0.039
(b) B1+PCF+PCF+PCF	0.917	0.916	0.954	0.026	0.914	0.899	0.931	0.042	0.941	0.917	0.950	0.032
(c) B1+SCR+SCR+SCR	0.918	0.916	0.951	0.026	0.913	0.904	0.940	0.037	0.939	0.918	0.951	0.032
(d) B1+DCA+DCA+DCA	0.920	0.914	0.951	0.027	0.912	0.896	0.932	0.043	0.931	0.909	0.946	0.036
(e) B1+DAM+DAM+DAM	0.919	0.916	0.954	0.025	0.919	0.905	0.938	0.039	0.936	0.916	0.950	0.032
(f) B1+CRM+CRM+CRM	0.917	0.915	0.953	0.026	0.912	0.899	0.936	0.040	0.931	0.912	0.945	0.035
(g) B1+CIM+CIM+CIM	0.917	0.913	0.953	0.026	0.910	0.896	0.934	0.041	0.935	0.911	0.945	0.035
(h) B1+RDE+DSE+DSE	0.919	0.916	0.953	0.025	0.921	0.905	0.938	0.039	0.937	0.916	0.948	0.032
(i) B1+PCF+SCR+DCA (SAFNet)	0.934	0.930	0.965	0.020	0.929	0.915	0.947	0.033	0.950	0.931	0.962	0.026

Effectiveness of Each Main Component in SAFNet. The experimental results of ablation studies on the effectiveness of the main components of our method are shown in Tab. 2. We design a set of baseline networks as comparisons. The RGB-D Baseline (B) in Tab. 2 (a) is the baseline network that takes both RGB and depth images as input. We employ a simple concatenation and convolution operation to integrate the RGB-D cross-modal features in the shallow, middle, and deep scales. The same simple fusion method is also used to replace the GFM in the saliency decoder to combine the features in different scales. Based on the baseline network, we successively replace the fusion strategies in different scales with the proposed PCF, SCR, and DCA. The quantitative performance is shown in Tab. 2 (b)-(d). In Tab. 2 (e) and (f), the GFM is adopted in the saliency decoder. Especially in Tab. 2 (e), the instance normalization operation in the PCF module is replaced by the batch normalization operation. Tab. 2 (f) shows the final performance of the proposed SAFNet.

Comparing the performances in Tab. 2 (a)-(d), we can observe that our proposed fusion strategies, PCF for the shallower scale, SCR for the middle scale, and DCA for the deep scale, can improve the performance of the baseline methods, which verifies the effectiveness of the proposed fusion methods. The improvement in Tab. 2 (f) over (d) indicates the validity of the GFM in the saliency decoder. We also validate the effectiveness of the 3D-2D Normalization module in the PCF module. Comparing Tab. 2 (e) and (f), we can see that the 3D-2D Normalization module slightly boosts the accuracy of the saliency modal. Theoretically, the 3D-2D Normalization operation projects the feature in 3D space to the 2D space. It ensures the distribution consistency of multi-scale features, which is necessary for the GFM in the saliency decoder.

Effectiveness of Fusion Strategies. In this section, we present a series of experiments to prove the superiority of the proposed fusion methods and the scale adaptive fusion strategy. For this purpose, we design the RGB-D Baseline (B1) by replacing the fusion modules, PCF, SCR, and DCA, in the proposed SAFNet with the simple concatenation and convolution operation (the GFM in the decoder remains). The performance is shown in Tab. 3 (a). In Tab. 3 (b)-(g), we employ the same fusion modules on different scales. In Tab. 3 (b)-(d), we adopt the proposed PCF, SCR, and DCA, respectively. The fusion module DAM in work [43] is utilized in Tab. 3 (e). Different from the DCA which is only based on the channel-wise attention module, the DAM consists of both channel-wise attention and spatial-wise attention. The CRM in Tab. 3 (f) is the multi-modal integration method proposed in work [14], and the CIM in Tab. 3 (g) is employed in work [46]. In Tab. 3 (h), we implement a similar idea in work [40], which uses the RDE in shallower layers and the DSE in deeper layers to enforce the information transmission between the RGB and depth streams. Tab. 3 (i) is the performance of the proposed SAFNet which utilizes the PCF, SCR, and DCA on different scales.

From the quantitative experiments in Tab. 3, we can see that all fusion methods achieve improvement over the baseline method (Tab. 3 (a)). The networks which employ scale adaptive fusion methods (Tab. 3 (h) and (i)) gain more promotion. Compared with the model in Tab. 3 (h), our method gives deep insight into the attributes of different scales in the network, and adopts early and mid-level fusion strategies to further enhance the interaction of multi-modal features. As a result, the proposed SAFNet achieves superior performance against the fusion strategies of existing methods.

5 Conclusion

In this work, we propose a Scale Adaptive Fusion Network (SAFNet) which takes account of different attributes of multi-scale features in the network for RGB-D SOD. For the shallow scale, we propose a Point Cloud Fusion (PCF) method to integrate the RGB and depth features in the 3D space. For the middle scale, a Spatial Contrast Refinement (SCR) module is designed to explore the structural information of the scene. For the deep scale, we adopt the Depth-aware Channel Attention (DCA) module to combine the semantic cues from the RGB-D features. In our work, both early and mid-level fusion strategies are adopted to enforce the in-depth fusion of the multi-modalities. Extensive experiments show that our proposed SAFNet achieves significant performance against the state-of-the-art algorithms and also verify the effectiveness of the scale adaptive fusion strategy exploited by our model.

Acknowledgements This work is supported by the Ministry of Science and Technology of the People's Republic of China no. 2018AAA0102003, National Natural Science Foundation of China under Grant no. 62006037, and the Fundamental Research Funds for the Central Universities no. DUT22JC06.

References

1. Achanta, R., Hemami, S.S., Estrada, F.J., Süsstrunk, S.: Frequency-tuned salient region detection. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1597–1604 (2009)
2. Borji, A., Sihite, D.N., Itti, L.: Salient object detection: A benchmark. In: European Conference on Computer Vision. pp. 414–429 (2012)
3. Chen, S., Fu, Y.: Progressively guided alternate refinement network for rgb-d salient object detection. In: European Conference on Computer Vision. pp. 520–538 (2020)
4. Desingh, K., Krishna, K.M., Rajan, D., Jawahar, C.V.: Depth really matters: Improving visual salient region detection with depth. In: British Machine Vision Conference (2013)
5. Ding, Y., Liu, Z., Huang, M., Shi, R., Wang, X.: Depth-aware saliency detection using convolutional neural networks. Journal of Visual Communication and Image Representation **61**, 1–9 (2019)
6. Fan, D., Cheng, M., Liu, Y., Li, T., Borji, A.: Structure-measure: A new way to evaluate foreground maps. In: IEEE International Conference on Computer Vision. pp. 4558–4567 (2017)
7. Fan, D., Gong, C., Cao, Y., Ren, B., Cheng, M., Borji, A.: Enhanced-alignment measure for binary foreground map evaluation. In: International Joint Conference on Artificial Intelligence. pp. 698–704 (2018)
8. Fan, D., Lin, Z., Zhao, J., Liu, Y., Zhang, Z., Hou, Q., Zhu, M., Cheng, M.: Rethinking rgb-d salient object detection: Models, datasets, and large-scale benchmarks. IEEE Transactions on neural networks and learning systems **32**(5), 2075–2089 (2020)
9. Garcia-Garcia, A., Gomez-Donoso, F., Rodríguez, J.G., Orts-Escolano, S., Cazorla, M., López, J.A.: Pointnet: A 3d convolutional neural network for real-time object class recognition. In: International Joint Conference on Neural Networks. pp. 1578–1584 (2016)
10. Han, J., Chen, H., Liu, N., Yan, C., Li, X.: Cnns-based rgb-d saliency detection via cross-view transfer and multiview fusion. IEEE Transactions on Cybernatics **48**(11), 3171–3183 (2018)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
12. Hou, Q., Cheng, M., Hu, X., Borji, A., Tu, Z., Torr, P.H.S.: Deeply supervised salient object detection with short connections. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(4), 815–828 (2019)
13. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning. vol. 37, pp. 448–456 (2015)
14. Ji, W., Li, J., Yu, S., Zhang, M., Piao, Y., Yao, S., Bi, Q., Ma, K., Zheng, Y., Lu, H., Cheng, L.: Calibrated rgb-d salient object detection. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 9471–9481 (2021)
15. Jin, W., Xu, J., Han, Q., Zhang, Y., Cheng, M.: Cdnet: Complementary depth network for rgb-d salient object detection. IEEE Transactions on Image Processing **30**, 3376–3390 (2021)
16. Ju, R., Liu, Y., Ren, T., Ge, L., Wu, G.: Depth-aware salient object detection using anisotropic center-surround difference. Signal Processing: Image Communication **38**, 115–126 (2015)

17. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. pp. 1106–1114 (2012)
18. Lang, C., Nguyen, T.V., Katti, H., Yadati, K., Kankanhalli, M.S., Yan, S.: Depth matters: Influence of depth cues on visual saliency. In: European Conference on Computer Vision. pp. 101–115 (2012)
19. Li, G., Liu, Z., Chen, M., Bai, Z., Lin, W., Ling, H.: Hierarchical alternate interaction network for rgb-d salient object detection. IEEE Transactions on Image Processing **30**, 3528–3542 (2021)
20. Li, G., Liu, Z., Ye, L., Wang, Y., Ling, H.: Cross-modal weighting network for rgb-d salient object detection. In: European Conference on Computer Vision. pp. 665–681 (2020)
21. Li, N., Ye, J., Ji, Y., Ling, H., Yu, J.: Saliency detection on light field. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(8), 1605–1616 (2017)
22. Liu, N., Zhang, N., Han, J.: Learning selective self-mutual attention for rgb-d saliency detection. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 13753–13762 (2020)
23. Liu, S., Huang, D., Wang, Y.: Receptive field block net for accurate and fast object detection. In: European Conference on Computer Vision. pp. 404–419 (2018)
24. Liu, Z., Shi, S., Duan, Q., Zhang, W., Zhao, P.: Salient object detection for rgb-d image by single stream recurrent convolution neural network. Neurocomputing **363**, 46–57 (2019)
25. Mahadevan, V., Vasconcelos, N.: Saliency-based discriminant tracking. In: Computer Vision and Pattern Recognition. pp. 1007–1013 (2009)
26. Niu, Y., Geng, Y., Li, X., Liu, F.: Leveraging stereopsis for saliency analysis. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 454–461 (2012)
27. Pang, Y., Zhang, L., Zhao, X., Lu, H.: Hierarchical dynamic filtering network for rgb-d salient object detection. In: European Conference on Computer Vision. pp. 235–252 (2020)
28. Peng, H., Li, B., Xiong, W., Hu, W., Ji, R.: Rgb-d salient object detection: A benchmark and algorithms. In: European Conference on Computer Vision. pp. 92–109 (2014)
29. Piao, Y., Ji, W., Li, J., Zhang, M., Lu, H.: Depth-induced multi-scale recurrent attention network for saliency detection. In: IEEE International Conference on Computer Vision. pp. 7253–7262 (2019)
30. Piao, Y., Rong, Z., Zhang, M., Ren, W., Lu, H.: A2dele: Adaptive and attentive depth distiller for efficient rgb-d salient object detection. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 9057–9066 (2020)
31. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In: Advances in Neural Information Processing Systems. pp. 5099–5108 (2017)
32. Quo, J., Ren, T., Bei, J.: Salient object detection for RGB-D image via saliency evolution. In: IEEE International Conference on Multimedia and Expo. pp. 1–6 (2016)
33. Ren, Z., Gao, S., Chia, L., Tsang, I.W.: Region-based saliency detection and its application in object recognition. IEEE Transactions on Circuits and Systems for Video Technology **24**(5), 769–779 (2014)
34. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015)

35. Sun, P., Zhang, W., Wang, H., Li, S., Li, X.: Deep rgb-d saliency detection with depth-sensitive attention and automatic multi-modal fusion. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1407–1417 (2021)
36. Ulyanov, D., Vedaldi, A., Lempitsky, V.S.: Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016)
37. Wu, Z., Su, L., Huang, Q.: Cascaded partial decoder for fast and accurate salient object detection. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3907–3916 (2019)
38. Yan, S., Song, X., Yu, C.: Sdcnet: Size divide and conquer network for salient object detection. In: Asian Conference on Computer Vision. pp. 637–653 (2020)
39. Yang, S., Lin, W., Lin, G., Jiang, Q., Liu, Z.: Progressive self-guided loss for salient object detection. IEEE Transactions on Image Processing **30**, 8426–8438 (2021)
40. Zhang, C., Cong, R., Lin, Q., Ma, L., Li, F., Zhao, Y., Kwong, S.: Cross-modality discrepant interaction network for rgb-d salient object detection. In: ACM Multimedia. pp. 2094–2102 (2021)
41. Zhang, J., Fan, D., Dai, Y., Anwar, S., Saleh, F.S., Zhang, T., Barnes, N.: Uc-net: Uncertainty inspired rgb-d saliency detection via conditional variational autoencoders. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 8579–8588 (2020)
42. Zhang, J., Fan, D., Dai, Y., Yu, X., Zhong, Y., Barnes, N., Shao, L.: Rgb-d saliency detection via cascaded mutual information minimization. In: IEEE International Conference on Computer Vision. pp. 4318–4327 (2021)
43. Zhang, M., Sun, X., Liu, J., Xu, S., Piao, Y., Lu, H.: Asymmetric two-stream architecture for accurate rgb-d saliency detection. In: European Conference on Computer Vision. pp. 374–390 (2020)
44. Zhao, X., Pang, Y., Zhang, L., , Lu, H., Ruan, X.: Self-supervised pretraining for rgb-d salient object detection. In: Association for the Advancement of Artificial Intelligence. pp. 3463–3471 (2022)
45. Zhao, X., Zhang, L., Pang, Y., Lu, H., Zhang, L.: A single stream network for robust and real-time rgb-d salient object detection. In: European Conference on Computer Vision. pp. 646–662 (2020)
46. Zhou, T., Fu, H., Chen, G., Zhou, Y., Fan, D., Shao, L.: Specificity-preserving rgb-d saliency detection. In: IEEE International Conference on Computer Vision. pp. 4661–4671 (2021)
47. Zhu, C., Cai, X., Huang, K., Li, T.H., Li, G.: Pdnet: Prior-model guided depth-enhanced network for salient object detection. In: IEEE International Conference on Multimedia and Expo. pp. 199–204 (2019)
48. Zhu, J., Wu, J., Xu, Y., Chang, E.I., Tu, Z.: Unsupervised object class discovery via saliency-guided multiple class learning. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**(4), 862–875 (2015)